·基础理论与方法·

对应分析在生态学研究中的应用

李宝红　董时富　孙振球

【摘要】 **目的**　探讨对应分析在生态学研究中的应用。**方法**　用对应分析方法对中国部分城市食品摄入量与男性胃癌死亡之间的关系进行分析。**结果**　从对应分析的因子负荷图中可以看出男性胃癌死亡率存在地区性差异。**结论**　因子负荷图直观地表述了中国部分城市与男性胃癌死亡率存在地区性差异，南方男性居民对米特别是精制的米以及食盐摄入较多，面类、浅色蔬菜摄入量较少。这可能暗示着存在一些致癌因素。

【关键词】 胃癌；对应分析；生态学研究；死亡率

An application of correspondence analysis method in the study of disease etiology　*LI Bao-hong*, DONG Shi-fu, SUN Zhen-qiu. *School of Public Health, Central South University, Changsha* 410078, China

【**Abstract**】 **Objective**　To explore the application value of correspondence analysis in ecological study. **Methods**　We adopted correspondence analysis method to analyze the relationship between the amount of food intake in some cities in China and the male gastric carcinoma mortality. **Results**　According to scatter plots of row and column points, there were regional differences among the male gastric carcinoma mortality in different cities of China. **Conclusion**　The scatter plot of row and column points indicated directly that there were regional differences among the male gastric carcinoma mortality in different cities of China. Males from the Southern part of the country ate more rice and salt, less wheaten food and fewer light vegetables than those from the northern parts, suggesting that there might be some carcinogenic factors in some food stuff involved.

【**Key words**】 Gastric carcinoma; Correspondence analysis; Ecological study; Mortality

在进行科学研究时经常会遇到包含大量信息的资料，其中既有大量的定性指标又有大量的定量指标，那么对于这些繁杂的数据如何进行分析呢？我们可以把这些数据整理成列联表的形式，用对应分析方法进行处理，把多维数据降成低维数据，这样既可以分析行指标还可以分析列指标以及对二者同时进行分析。本文旨在介绍对应分析（correspondence analysis），并以该方法对中国部分城市食品摄入量与男性胃癌死亡之间的关系进行分析。

基本原理

对应分析又称相应分析[1-5]，主要用于分析二维数据矩阵中行因素和列因素间的关系。对应分析的基本原理是，对二维数据矩阵进行适当的变换（即对应变换），使变换后的数据对行与对列是相对应的，从而可以同时对行和对列进行分析，以便发现行列因素间的关系。实际上它是将R-型因子分析与Q-型因子分析相结合，对指标与样品同时进行分类的一种多元统计分析方法。

方法和步骤[6,7]：设有 n 个样品（或观测对象），每个样品观测 m 个变量，得原始资料的矩阵为 $X=(x_{ij})_{n\times m}$，式中 x_{ij} 为第 i 个样品，第 j 个观测变量的观测值。$i=1,2,\cdots,n$；$j=1,2,\cdots,m$。即：

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

1. 将 X 按行、列分别求和

$$T=\sum_{i=1}^{n}\sum_{j=1}^{m}x_{ij}$$

2. 对原始数据 x_{ij} 进行对应变换，得矩阵 $Z=(z_{ij})_{n\times m}$

式中：$z_{ij}=\dfrac{x_{ij}-x_{i\cdot}x_{\cdot j}/T}{\sqrt{x_{i\cdot}x_{\cdot j}}}(i=1,2,\cdots,n;j=1,2,\cdots,m)$

$$x_{i\cdot}=\sum_{j=1}^{m}x_{ij},\quad x_{\cdot j}=\sum_{i=1}^{n}x_{ij}$$

作者单位：410078 长沙，中南大学公共卫生学院（李宝红、孙振球）；华中科技大学同济医学院卫生统计教研室（董时富）

3. 进行 R-型因子分析。计算协方差矩阵 $Z'Z$ 的特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m$，其中 Z' 为 Z 的转置矩阵，根据累积贡献率≥75%，80%，85%等，取前 p 个特征值，计算相应的单位特征向量 $U_1, U_2, \cdots, U_p$，从而得到 R-型因子载荷矩阵，并在两两因子轴组成的平面上作出变量散点图。

4. 进行 Q-型因子分析。对上面计算出的 p 个特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$，计算矩阵 ZZ' 的单位特征向量，$V_1 = ZU_{1(m\times 1)}$，$V_2 = ZU_2, \cdots, V_p = ZU_p$，从而得到 Q-型因子载荷矩阵，并在与 R-型相应的因子平面上作出样品散点图。

5. 对对应分析的结果合理地进行解释与推断。根据对应分析图作如下的解释：相邻的变量点关系密切，相邻的样品点具有相似的性质，同一点群族的样品点将由同族邻近的变量点来表征。对应分析是把R-型和Q-型因子分析结合起来，把变量和样品同时反映到同一个因子轴的坐标系中对变量和样品一起进行分析。其基本思想就是将一个列联表的数据形式在较低维的空间中表示出来，起到降维的作用。作对应变换是对应分析的关键，其余的分析与因子分析类似，只是在因子的解释上，既可以对行因素（样品）及列因素（变量）单独进行分析又可以同时进行分析，这是对应分析的优点。

对应分析可用软件 SAS 6.04以上版本或 SPSS 4.0以上版本进行分析。

实例分析

胃癌是常见的恶性肿瘤之一。据国际癌症研究机构（IARC，1982 年）公布的 105 个国家和地区的资料表明，世界胃癌的年发病率平均为17.6/10 万，占各部位恶性肿瘤的第五位。我国胃癌死亡率与世界其他国家比较，处于较高水平，调整死亡率为 15.41/10 万（男性为 20.93/10 万，女性为 10.16/10 万）[8]。近几十年来世界各国都在进行胃癌的病因学研究，无论是从生物学的角度还是流行病学的角度都在不断地深入。但到目前为止，胃癌的病因尚未确定。目前多数学者认为环境因素尤其是膳食因素是胃癌发病的主要原因。因此继续从事食物因素与癌症关系的分析，应有现实意义。

本文采用 1982 年部分城市男性居民胃癌粗死亡率与对应这些城市男性居民的部分食品摄入量之间进行对应分析，意在通过此寻找致胃癌的食物危险因素和保护因素，并同时介绍对应分析方法在医学上的应用。本文引用姜又红等[9]的二维表（表 1）。其中该文献的部分城市男性平均每人每日各类食品摄入量资料引自原中国预防医学中心卫生研究所“1982 年全国营养调查总结”，部分城市男性胃癌粗死亡率资料引自我国“1982 年全国卫生统计年报资料”。

表1 我国 8 城市男性胃癌死亡率（/10 万）与各类食品摄入量（g）

变量	城市							
	北京（Y_1）	天津（Y_2）	沈阳（Y_3）	长春（Y_4）	上海（Y_5）	南京（Y_6）	苏州（Y_7）	西安（Y_8）
X_1（死亡率）	0.013 40	0.015 01	0.015 15	0.014 63	0.022 39	0.021 64	0.022 60	0.015 04
X_2（大米）	136.0	177.0	195.0	181.0	380.0	359.0	350.0	92.0
X_3（面粉）	289.0	318.5	208.0	261.0	53.3	48.0	71.0	404.0
X_4（薯类）	54.0	70.0	166.0	4.0	34.0	57.0	47.0	41.0
X_5（干豆类）	3.8	11.6	1.1	20.1	5.0	8.4	11.2	3.9
X_6（豆制品）	2.5	2.2	19.2	18.8	10.8	25.4	15.0	10.7
X_7（浅色蔬菜）	492.3	568.5	255.0	410.0	139.5	193.0	287.0	268.0
X_8（绿色蔬菜）	50.3	14.0	48.0	70.0	135.5	167.0	90.0	67.0
X_9（咸菜）	3.2	0.7	8.1	1.2	8.6	7.4	11.6	8.4
X_{10}（水果）	154.7	165.1	8.8	18.2	15.9	22.1	19.3	51.8
X_{11}（乳类）	61.4	4.7	21.0	18.0	12.0	3.3	8.5	17.7
X_{12}（蛋类）	24.1	37.9	4.0	90.8	23.6	18.0	17.8	19.4
X_{13}（猪肉）	46.2	44.1	31.0	59.7	54.7	57.4	56.9	34.8
X_{14}（鱼虾类）	10.5	9.9	6.0	11.5	50.7	38.9	15.8	1.1
X_{15}（淀粉类）	13.5	8.6	4.1	22.4	22.2	12.9	20.9	11.0
X_{16}（植物油）	21.6	21.0	7.9	18.2	7.1	2.9	1.7	5.5
X_{17}（食盐）	12.7	10.3	7.0	9.2	13.6	10.0	14.0	15.0

表2　惯性贡献

维数	特征值	惯性	χ^2 统计量	显著性检验	惯性贡献率		特征值标准差及因子间特征值相关系数	
					各因子贡献率	累积贡献率	标准差	相关系数
1	0.435 57	0.189 72			0.604 42	0.604	0.009	0.034
2	0.225 64	0.050 91			0.162 20	0.767	0.012	
3	0.188 41	0.035 50			0.113 10	0.880		
4	0.139 01	0.019 32			0.061 56	0.941		
5	0.098 11	0.009 63			0.030 67	0.972		
6	0.072 99	0.005 33			0.016 97	0.989		
7	0.058 96	0.003 48			0.011 08	1.000		
合计		0.313 89	2 865.987	0.000[a]	1.000 00	1.000		

注：[a]自由度为 112

把表 1 中 8 个城市作为样品（$Y_j, j=1,2,\cdots,8$），1 个胃癌死亡率 16 个各类食品摄入量作为变量（$X_i, i=1,2,\cdots,17$），对数据矩阵进行对应分析，用 SPSS 13.0软件完成，结果见表 2 和图 1（为对应分析中最重要的结果输出部分）。

从表 2 可知，惯性等于各因子特征值的平方，如 $0.189\ 72=0.435\ 57^2$ 用以表示各因子的重要性；χ^2 统计量也就是表 1 的 χ^2 检验，自由度为 112，$P=0.000$。根据表 2 可见，前二个特征值的累积贡献率为0.767，即76.7%，表明这二个维度（dimension）能够解释总信息量的76.7%，由前二个因子轴构成的二维投影图（图 1）已包含了原向量信息的76.7%。

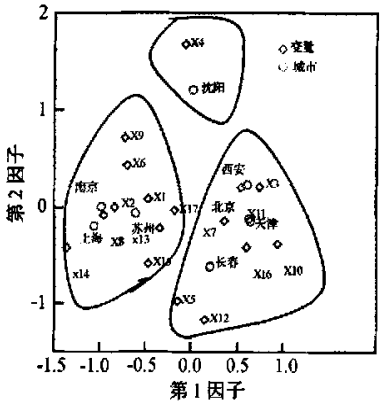

图1　对应分析因子负荷图

由图 1 可以看出，对应分析在同一平面上给出了样品点群与变量点群，从而可以直观地研究样品点群与变量点群之间的关系，这是对应分析的主要特点。依据各样品点和变量点的簇集情况，可将分成 3 个点群：

Ⅰ城市：上海（Y_5）、南京（Y_6）、苏州（Y_7）

变量：大米（X_2）、豆制品（X_6）、绿色蔬菜（X_8）、咸菜（X_9）、猪肉（X_{13}）、鱼虾类（X_{14}）、淀粉类（X_{15}）、食盐（X_{17}）

Ⅱ城市：北京（Y_1）、天津（Y_2）、长春（Y_4）、西安（Y_8）

变量：面粉（X_3）、干豆类（X_5）、浅色蔬菜（X_7）、水果（X_{10}）、乳类（X_{11}）、蛋类（X_{12}）、植物油（X_{16}）

Ⅲ城市：沈阳（Y_3）

变量：薯类（X_4）

阅读对应分析图的原则是：代表行变量某个类别的点，与代表列变量某个类别的点距离较近，则二者有较强的关联性；若距离较远，则表明二者关联性较弱或无关联性。根据对应分析图可作如下解释：由图 1 可见，点群族Ⅰ城市中上海（Y_5）、南京（Y_6）、苏州（Y_7）为胃癌高死亡率地区，而点群族Ⅰ城市中的变量为大米（X_2）、豆制品（X_6）、绿色蔬菜（X_8）、咸菜（X_9）、猪肉（X_{13}）、鱼虾类（X_{14}）、淀粉类（X_{15}）、食盐（X_{17}）靠近高胃癌死亡率地区数据群，说明这些变量与胃癌死亡率呈正相关。例如，大米与胃癌有较大的关联性，此结果和一些学者的报道相一致[10]，原因是由于精制米类缺少微量元素，且由于米类摄入的增加，从而相对地减少了其他食物的摄入，因此使患胃癌的危险性相对增加，如上海、南京、苏州的男性居民主食以米类为主，胃癌死亡率明显高于其他城市。尽管有报道证明绿色蔬菜和肉类摄入量的适当增加有防胃癌的作用[11]，但从本次研究的结果来看，并没有因为绿色蔬菜和猪肉摄入的增加而消弱米类对胃癌的作用。咸菜和鱼虾类以及食盐与胃癌死亡率呈正相关，此结果和其他学者报道的日本等国家的情况一致[12]，可能由于咸菜中富含硝酸盐、亚硝酸盐，促使亚硝胺的生成，而亚硝胺是一个典型的致癌物。点群族Ⅱ城市中北京（Y_1）、天津（Y_2）、长春（Y_4）、西安（Y_8）为胃癌低死

亡率地区，而点群族Ⅱ城市中的变量为面粉（X_3）、干豆类（X_5）、浅色蔬菜（X_7）、水果（X_{10}）、乳类（X_{11}）、蛋类（X_{12}）、植物油（X_{16}）靠近低胃癌死亡率地区数据群，这些变量与胃癌死亡率呈负相关，说明这些变量对患胃癌的危险性可能相对减少，也许可作为保护性因素。如面类（X_3）、浅色蔬菜（X_7）靠近低胃癌死亡率数据群，说明与胃癌呈负的关联性。由于此类食物中富含维生素，而维生素对肿瘤的形成有直接的抑制作用。植物油亦有防癌的作用。由于植物油既有利于对胃黏膜的保护，又可促进胃黏膜修复，因此可间接地起到防癌的作用。点群族Ⅲ城市中沈阳（Y_3）介于前两者之间，为中等胃癌死亡率地区，而变量薯类（X_4）靠近它，因此可认为它对胃癌的发生没有明显关联。

从对应分析的结果可以看出，人们的饮食，特别是南方饮食习惯的男性居民，应适当减少米类特别是精制米类的摄入，而增加面类、植物油的摄入，多吃清淡的食物，减少食盐的摄入量，从膳食角度预防胃癌的发生，降低胃癌的死亡率。

讨　论

对应分析最早用于处理二维列联表资料，一般要求数据不小于 0，在有数据小于 0 时，将每个数据加上同一适当的正常数即可，不影响对结果的分析。分析中主要取前两个主成分，这就要求二者的累计贡献率较大，一般以大于 75% 为宜。对应分析对小概率事件较为敏感，通过因子负荷图，能发现各变量与每个样本的亲密程度，每一样本点或样本点群，可以用靠近它们的变量点来描述，所以这对提示个性尤其是有意义的小概率事件十分有用[6,13]。

对应分析具备一般主成分分析的特征，即降维作用。虽然要分析的变量较多，但真正说明问题的只是方差贡献最大的少数几个因子。这就大大减少了工作量，起到降维的作用，而且各因子彼此正交，不存在多重共线问题。在医学研究中，观测变量往往很多，定性、定量的指标都有，彼此又往往高度相关，给数据分析带来许多难题。若事先用对应分析进行预处理，仅取方差贡献最大的前几个因子参加下一步的分析，将带来极大的方便。

综上所述，可以看出，对应分析在医学研究中有很好的应用前景。

参 考 文 献

[1] Ludovic L, Alain M, Kenneth W. Multivariate descriptive statistical analysis—correspondence analysis and related techniques for large matrices. Translated by Elisabeth Moraillon Berry, etc. New York, 1984: 30-62.

[2] Hill MO. Correspondence analysis. Encyclopedia of Statistical Sciences. New York, 1982: 204-210.

[3] 裴鑫德. 多元统计分析及其应用. 1 版. 北京：北京农业大学出版社，1991：550-563.

[4] 张明立，于秀林. 多元统计分析方法及程序——在体育科学中的应用. 1 版. 北京：北京体育学院出版社，1991：187-201.

[5] 陈峰，杨树勤. 相应分析及其在多种疾病聚集性分析中的应用. 中国卫生统计，1999，16(2)：114-117.

[6] 郑金平，王琳娜，张晋昕. 对应分析法对科技成果未获奖因素再分析. 现代预防医学，1999，26(4)：473-475.

[7] 梁荣辉，张科，刘若群. 对应分析方法在学生体质调研样品分类中的应用. 中国卫生统计，1999，16(1)：98-99.

[8] 周利峰. 胃癌流行病学研究近况. 实用预防医学，1997，14(2)：124.

[9] 姜又红，杨军，鞠振宇，等. 中国部分城市食品摄入量与男性胃癌死亡的多变量分析. 中国公共卫生，2000，16(5)：404-405.

[10] Gonzalez A. Nutritional factors and gastric cancer in Spai. Am J Epidemiol, 1994, 139: 466-473.

[11] Ramon JM, Serra L, Cerdo C, et al. Dietary factors and gastric cancer risk: a case-control study in Spain. Cancer, 1993, 71: 1731-1735.

[12] Tuyns AJ, Kaaks R, Haelterman M, et al. Diet and gastric cancer. A case - control Study in Belgium. Int J Cancer, 1992, 51: 1-6.

[13] 刘韵源，郭万德. 对应分析及其在医学中的应用(二). 中华预防医学杂志，1985，19(3)：175-177.

（收稿日期：2006-12-25）

（本文编辑：张林东）